University of Málaga

Health Engineering

Laboratory Task

Clustering

Author

Alejandro Domínguez Recio

Course

Intelligent Systems

Teachers

Enrique Domínguez Merino

Jesús de Benito Picazo

Introduction

In this practice we going to evaluate a bioinformatics dataset with the clustering classifier K-means. We are going to evaluate the results depending on the distance measurement method, euclidean or manhattan, the mean square error or the number of predefined clusters . We are also going to detail the general characteristics of the dataset

How are we going to do it?

To do the different studies we will have the following structure in each of then.

- ➔ **Dataset context**

In this part we going to write a little introduction of the dataset context commenting on data type and prediction target.

- ➔ **Number of classes**

We going to describe the number of classes. To do that we going to observe the atributtes weka panel and we going to select the class attribute.

- ➔ **Number of attributes**

For the number of attributes we going to inspect the weka attributes panel or open the file in text mode and inspect the characteristics.

- ➔ **Number of samples or instances**

To know the number of samples we can proceed as in the previous section by inspecting the weka panels or opening the document in text mode and seeing its characteristics.

➔ **Values of permance measures**

Cluster error

We can obtain the cluster error directly from the classifier out. This measure shows us the sum of the distance of the samples towards their centroids in each iteration

Incorrectly clustered instances

This measure shows us the number of incorrectly classified instances of our test set.

Number of iterations

This measure shows us the number of iterations of the kmeans algorithm to complete the clustering. This measure depends on how classified the data is initially as the number of target clusters.

➔ **Evaluation of the results**

In this section we are goint to compare the results of the different performance measures.

➔ **Last conclusions**

Finally, we are going explain possible reasons why some test have better results than others.

➔ Introduction

The main idea of this clustering problem is given a bioinformatics dataset with a number of genes group these and determine if they belong to a type of cancer or another and evaluate its performance.

In this dataset we have **144** instances and each of them has **16064** attributes, one of them is the class attribute.

➔Number of classes

In this case our class attribute can take **14** ***values***.

➔ Number of attributes

In this case we have **16064** which determine the genes studied with a possible relationship to cancer.

➔ Number of samples

In this case in particular the number of samples or instances is **144.**

➔Classification of data in preprocess tab

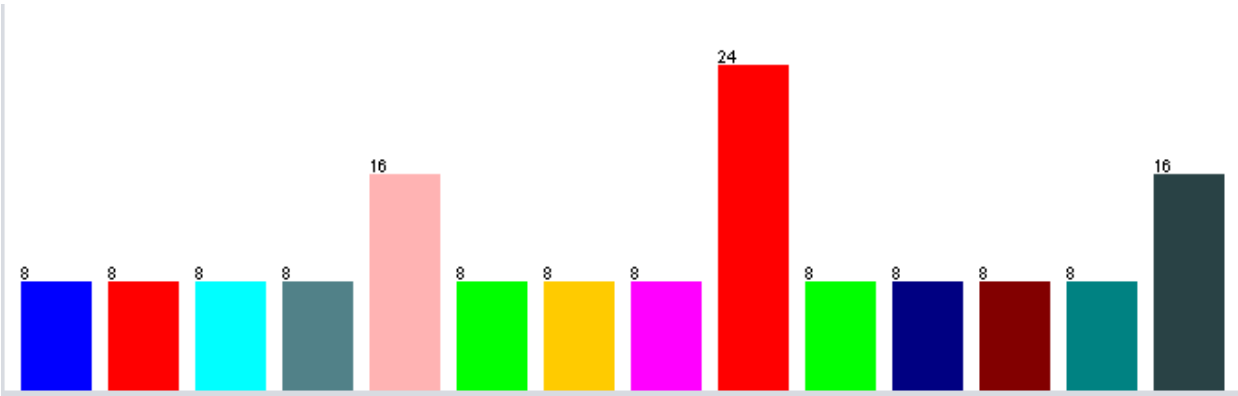

Ilustración 1: ClassificationDataPreprocessTab

EUCLIDEAN DISTANCE TESTS

- Number of cluster : **14**

➔ **Values of permance measures**

Cluster error : **27175.212505809**

Number of iterations: **8**

Incorrectly clustered instances: **95.0 – 65.9722 %**

Clustered instances and preprocess data:

	Cluster#													
Full Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13
(144.0)	(24.0)	(7.0)	(19.0)	(16.0)	(17.0)	(7.0)	(10.0)	(5.0)	(3.0)	(8.0)	(1.0)	(5.0)	(13.0)	(9.0)

As we can see, the proportion of instances per cluster does not correspond to the initial ones, denoting that even though the number of clusters chosen is the same as the number of classes in the k-means dataset, it reorganizes them in different proportions.

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Lung

Cluster 2 <-- Melanoma

Cluster 3 <-- Lymphoma

Cluster 4 <-- CNS

Cluster 5 <-- Renal

Cluster 6 <-- Leukemia

Cluster 7 <-- No class

Cluster 8 <-- No class

Cluster 9 <-- Colorectal

Cluster 10 <-- Ovary

Cluster 11 <-- No class

Cluster 12 <-- Bladder

Cluster 13 <-- Prostate

In the case of the number of classes associated with clusters, we can see that certain clusters do not have any class assigned and therefore without the possibility of being classified.

- Number of cluster : **20**

➔ **Values of permance measures**

Cluster error : **24804.83945389218**

Number of iterations: **7**

Incorrectly clustered instances: 84**.0 – 58.3333 %**

Clustered instances and preprocess data:

	Cluster#														
Full Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
(144.0)	(18.0)	(3.0)	(14.0)	(9.0)	(6.0)	(5.0)	(10.0)	(2.0)	(2.0)	(12.0)	(1.0)	(5.0)	(10.0)	(4.0)	
	14	15	16	17	18	19									
	[5.0]	[5.0]	[15.0]	[11.0]	[3.0]	[4.0]									

Obviously, the proportion of instances per cluster does not correspond to the initial ones, because the number of chosen cluster is greater than the initial number of classes. Although we can say from the proportion obtained that it does not have any cluster with a disproportionate number of instances compared to the initial one. Range between [2.0 18.0].

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Melanoma

Cluster 2 <-- Pancreas

Cluster 3 <-- No class

Cluster 4 <-- CNS

Cluster 5 <-- Ovary

Cluster 6 <-- Leukemia

Cluster 7 <-- No class

Cluster 8 <-- Colorectal

Cluster 9 <-- Bladder

Cluster 10 <-- No class

Cluster 11 <-- No class

Cluster 12 <-- Lung

Cluster 13 <-- No class

Cluster 14 <-- No class

Cluster 15 <-- Prostate

Cluster 16 <-- Lymphoma

Cluster 17 <-- Mesothelioma

Cluster 18 <-- Uterus__Adeno

Cluster 19 <-- Renal

Compared to the previous test in which we chose the same number of clusters as the classes the dataset has and it returned a classification in which we had classes without an assigned cluster, in this case each class has a cluster associated, however we also have clusters without class for the excess of seven with respect to the initials

- Number of cluster : **2**

➔ **Values of permance measures**

Cluster error : **44018.2641025381**

Number of iterations: **6**

Incorrectly clustered instances: 112**.0 – 77.7778 %**

Clustered instances and preprocess data:

	Cluster#	
Full Data	0	1
(144.0)	(83.0)	(61.0)

In this case we have only two clusters with a similar proportion of instances in each one. The objective of choosing this very low cluster number is the comparison between a choice of the cluster number that is smaller and greater than the initial cluster number, thus seeing the effects of this on the results.

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Leukemia

The choice of classes associated with the clusters is given by the greatest number of these among the different initial clusters. For this reason, Breast and Leukemia have been chosen since these in comparison with the rest of the classes are the ones that have the highest proportion in the different clusters.

MANHATTAN DISTANCE TESTS

- Number of cluster : **14**

➔ **Values of permance measures**

Cluster error : **156931.0968644462**

Number of iterations: **11**

Incorrectly clustered instances: **92.0 – 63.8889 %**

Clustered instances and preprocess data:

	Cluster#													
Full Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13
(144.0)	(19.0)	(6.0)	(17.0)	(16.0)	(14.0)	(7.0)	(11.0)	(9.0)	(7.0)	(11.0)	(2.0)	(5.0)	(14.0)	(6.0)

As we can see, the proportion of instances per cluster does not correspond to the initial ones, denoting that even though the number of clusters chosen is the same as the number of classes in the k-means dataset, it reorganizes them in different proportions. Range between [2.0 19.0]

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Melanoma

Cluster 2 <-- Uterus__Adeno

Cluster 3 <-- Lymphoma

Cluster 4 <-- CNS

Cluster 5 <-- Renal

Cluster 6 <-- Leukemia

Cluster 7 <-- Prostate

Cluster 8 <-- Colorectal

Cluster 9 <-- Bladder

Cluster 10 <-- Ovary

Cluster 11 <-- No class

Cluster 12 <-- Lung

Cluster 13 <-- Mesothelioma

In the case of the number of classes associated with clusters, we can see that only one cluster has not class assigned.

- Number of cluster : **20**

➔ **Values of permance measures**

Cluster error : **144231.10152121756**

Number of iterations: **7**

Incorrectly clustered instances: 84**.0 – 58.3333 %**

Clustered instances and preprocess data:

	Cluster#														
Full Data	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
(144.0)	(15.0)	(6.0)	(10.0)	(10.0)	(8.0)	(7.0)	(10.0)	(3.0)	(2.0)	(6.0)	(1.0)	(5.0)	(11.0)	(3.0)	
	14	15	16	17	18	19									
	[7.0]	[4.0]	[13.0]	[9.0]	[7.0]	[7.0]									

Obviously, the proportion of instances per cluster does not correspond to the initial ones, because the number of chosen cluster is greater than the initial number of classes. Although we can say from the proportion obtained that it does not have any cluster with a disproportionate number of instances compared to the initial one. Range between [1.0 15.0].

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Melanoma

Cluster 2 <-- Pancreas

Cluster 3 <-- No class

Cluster 4 <-- No class

Cluster 5 <-- Bladder

Cluster 6 <-- Leukemia

Cluster 7 <-- No class

Cluster 8 <-- No class

Cluster 9 <-- Colorectal

Cluster 10 <-- No class

Cluster 11 <-- No class

Cluster 12 <-- Lung

Cluster 13 <-- Ovary

Cluster 14 <-- CNS

Cluster 15 <-- Prostate

Cluster 16 <-- Lymphoma

Cluster 17 <-- Mesothelioma

Cluster 18 <-- Uterus__Adeno

Cluster 19 <-- Renal

Compared to the previous test in which we chose the same number of clusters as the classes the dataset has and it returned a classification in which we had classes without an assigned cluster, in this case each class has a cluster associated, however we also have clusters without class for the excess of seven with respect to the initials. Also the association of cluster and class has changed with respect to the test with Euclidean distance

- Number of cluster : **2**

➔ **Values of permance measures**

Cluster error : **198782.6685991927**

Number of iterations: **6**

Incorrectly clustered instances: 112**.0 – 77.7778 %**

Clustered instances and preprocess data:

	Cluster#	
Full Data	0	1
(144.0)	(78.0)	(66.0)

In this case we have only two clusters with a similar proportion of instances in each one. There are almost no differences with respect to the test with Euclidean distances in the proportions of clustered instances, it is more coincide in the number of incorrectly clustered instances.

Classes and clusters:

Cluster 0 <-- Breast

Cluster 1 <-- Leukemia

The choice of classes associated with the clusters is given by the greatest number of these among the different initial clusters. For this reason, Breast and Leukemia have been chosen since these in comparison with the rest of the classes are the ones that have the highest proportion in the different clusters. There are no differences in this aspect with respect to the test with Euclidean distance

➔ **Evaluation of the results**

Test	Number Clusters	Incorrectly Clustered Instances	SumOf Squared Errors	NumberOf Iterations	Classes/Clusters
Euclidean					
Test14	14	95	27175.2125	8	11/14
Test20	20	84	24804.8394	7	14/20
Test2	2	112	44018.2641	6	2/2
Manhattan					
Test14	14	92	156931.0968	11	13/14
Test20	20	84	144231.1015	7	14/20
Test2	2	112	198782.6685	6	2/2

➔ **Last conclusions**

By way of conclusion, we can say that clustering with the Manhattan distance measurement method provides us with better values both in the number of incorrectly clustered instances, although in the tests with 20 and 2 predefined clusters it provides us with the same results, in the quadratic sum of errors, regardless of the number of predefined clusters and the number of clustered classes. In turn, we add that better results are observed in proportion to the number of initially predefined clusters.